DOCUMENT RESUME

ED 306 302 TM 013 193

AUTHOR Schattgen, Sharon; And Others
TITLE A Validity-Based Model for the Evaluation of a Criterion-Referenced Test.
PUB DATE Mar 89
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Content Validity; *Criterion Referenced Tests; Elementary School Students; Elementary Secondary Education; Grade 3; Grade 6; Grade 8; Grade 10; Mastery Tests; Models; Norm Referenced Tests; School Districts; Secondary School Students; State Programs; *State Standards; Testing Programs; *Test Validity
IDENTIFIERS Approval Needs; Test Batteries

ABSTRACT

This paper describes a model for the evaluation and approval of a test battery for compliance with a midwestern state law mandating criterion-referenced testing of specific objectives. Standards specifying that the test scores must demonstrate content validity and criterion-related validity form the foundation of the model. The model also incorporates a three-stage approval process that promotes communication between the district applying for test approval and the state education agency. A district seeking approval of an norm-referenced battery administered its tests and the state tests concurrently to representative samples of students in grades 3, 6, 8, and 10. Results were submitted to Pearson product-moment coefficients, phi coefficients, and one-way analysis of variance procedures. District and state scores were clearly comparable at the overall subject level for three of the four grades. With respect to objective mastery rates, the degree of comparability was low across all grades and subjects. In general, the results support the effectiveness of the validity-based model. Five data tables are included. Criteria for approval of district-developed or published tests for use in measuring key skills, and Parts I and II of the Application for Test Approval are appended. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED306302

A VALIDITY-BASED MODEL FOR THE EVALUATION OF A CRITERION-REFERENCED TEST

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

S. F. SCHATTGEN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Sharon Schattgen
University of Missouri

Otis Baker
Missouri Department of Elementary
and Secondary Education

Steven Osterlind
University of Missouri

James Friedebach
Missouri Department of Elementary
and Secondary Education

Charles Oviatt
Missouri Department of Elementary
and Secondary Education

Paper presented at the Annual Meeting of the
American Educational Research Association

San Francisco, California, March 1989

TM013193

BEST COPY AVAILABLE

Abstract

This paper describes a validity-based model for the evaluation and approval of a test battery for compliance with a state law mandating criterion-referenced testing of specific objectives. Standards specifying that the test scores must demonstrate content and criterion-related validity form the foundation of the model. The model also incorporates a three-stage approval process which promotes communication between the district applying for test approval and the state education agency. Evidence for the criterion-related validity of a battery considered for approval, obtained during implementation of the model and using test scores yielded by the state-developed test as the criterion, is reported.

A Validity-Based Model for the Evaluation of a Criterion-Referenced Test

Legislation passed in a midwestern state as a part of the educational reform movement mandated that each school district periodically test students' mastery of certain objectives in seven subjects: reading, language arts, English, mathematics, science, social studies, and civics. Approximately 100 objectives per level for grades 2-10 were identified for paper-and-pencil testing by the state education agency. State-developed, criterion-referenced tests are available for measurement of these objectives. A local school district may, however, choose to construct its own tests or contract with a publisher to develop or adapt a commercial test. According to the legislation, an adapted or newly developed test battery must be evaluated and approved by the state education agency before a district can utilize it.

The state education agency developed and the state board of education adopted a validity-based model for such an evaluation and approval process. The model includes criteria which ensure that the tests approved for use yield scores that demonstrate important psychometric characteristics, especially content and criterion-related validity. The approval process is designed to be effective and efficient, both from the perspective of the applicant and the state agency. While at least one other state education agency has established procedures for test approval (Madaus, Airasian, Hambleton, Consalvo, Orlandi, 1979), this model represents a simpler and more manageable alternative.

The purpose of this paper is to describe in detail the validity-based model developed by the state education agency for the evaluation of a criterion-referenced test. Data representing the comparability of the state battery to a specific district battery are presented to illustrate how evidence for the criterion-referenced validity of district scores is obtained.

Approval Criteria

The criteria for approval of district tests (see Appendix A) are based on the Standards for Educational and Psychological Testing

(American Psychological Association, 1985) and on guidelines for evaluating criterion-referenced tests (Hambleton and Eignor, 1978). Other sources used to develop the criteria include accepted procedures for the construction of criterion-referenced tests (Berk, 1980, 1984) as well as specific provisions of the testing legislation.

The criteria are organized in accordance with basic psychometric concepts: (a) rationale and plan for test development, (b) content and criterion-related validity of scores, (c) reliability of scores and mastery classifications, (d) reporting and interpretation of results, and (e) test and item bias. There are specific standards corresponding to each concept that district tests must meet in order to be approved. In particular, a district constructing or adapting a test battery must submit to the state education agency a written plan that outlines the steps to be taken prior to embarking upon the development or adaptation process. Standards pertaining to test length, score and mastery-classification reliability, and test and item bias are also included. In addition, there are particular requirements pertaining to score reporting for individual students and groups of students and for providing aids for interpretation of test results.

The validity section is the cornerstone of the criteria. It specifies that the district battery must measure the entire set of state objectives for the grades selected for testing. In addition, a content specification for each objective must be used in the test development/adaptation process, and evidence of content validity in the form of item-objective congruence ratings from content experts must be obtained. Finally, a concurrent administration of the district battery and the state battery is required in order to establish the comparability of the two instruments. This comparability study allows for the collection of criterion-related validity evidence for the district tests using the state tests as the criterion.

Three-Stage Evaluation and Approval Process

A three-stage process was established to determine whether district tests meet the approval criteria. Each stage involves different but related activities and products. The initial stage occurs

before the tests are developed/adapted when the district files Part I of the "Application for Test Approval" (see Appendix B). The application must always be submitted by a district, even if a publisher is developing or adapting a battery. At this stage, the plan and rationale are submitted to the state agency for review.

The second stage involves completing and submitting Part II of the "Application for Test Approval" (see Appendix C) after psychometric data have been obtained. This part requires the district to indicate whether specific steps have been taken and certain data have been collected.

The third and final stage occurs when state education agency personnel review and verify the data. The criteria are applied to the products from the first and second stage activities, resulting in an ongoing evaluation of the development/adaptation process, the tests, and the psychometric data.

Because each stage builds upon the one preceding it, a cumulative effect results with respect to gaining approval as progression through the stages occurs. This process also fosters communication between the district and the state education agency.

Special worksheets and data forms accompany the applications for test approval. The worksheets were designed to facilitate completion of the application, while the data forms allow test information to be efficiently recorded and reviewed.

Comparability Study

To comply with the comparability-study criterion, a district seeking approval of an adapted norm-referenced battery administered its tests and the state tests concurrently to representative samples of students in grades 3, 6, 8, and 10. The sample sizes for each grade and subject were adequate and are presented in Table 1.

[Insert Table 1 about here]

Three types of statistics were computed on the data yielded by these concurrent administrations in order to investigate the comparability of the two batteries. Pearson product-moment coefficients

were computed on corresponding subject scores in order to investigate the criterion-related validity of the district's scores. Additional validity data were obtained by computing phi coefficients to determine the degree of relationship between corresponding objective mastery rates for the two batteries. In order to determine whether there were statistically significant differences between the proportions of students mastering each objective, one-way analysis of variance procedures were applied to two sets of mastery rates for the two batteries. The district's mastery standards were applied to the district data in the first set, while the state's mastery standards were applied to the district data (resulting in identical mastery standards being applied to both batteries) in the second set.

Results

Table 2 shows the Pearson product-moment coefficients for corresponding subject test scores for the two batteries.

[Insert Table 2 about here]

These indices of criterion-related validity for the district tests were strong for all subject tests in grades 3, 6, and 8 but were moderate for the subjects in grade 10.

Summary statistics for the distributions of phi coefficients (indices of consistency of mastery classifications yielded by the two batteries) are shown in Table 3. These coefficients were computed on district data which reflected district mastery standards; these standards varied across objectives and were not always identical to the state standards.

[Insert Table 3 about here]

The median, low, and high values were similar across grades and subjects and suggested that there was only a weak degree of comparability between the two batteries with respect to classifying students as masters or nonmasters of objectives.

The data in Tables 4 and 5, complementary to the data in Table 3, show the results of applying one-way analysis of variance procedures to

the proportions mastering each objective on the two batteries. Table 4 shows the percentage of objectives with significantly different mastery rates when the district's mastery standards were applied to district scores. Table 5 shows the percentage of objectives with significantly different mastery rates when identical standards (the state's) were applied to district scores. An alpha level of .05 represents significance on both tables.

[Insert Tables 4 and 5 about here]

Across grades and subjects, the percentage of objectives having significantly different mastery rates was high for both conditions. There were far more objectives with significantly different mastery rates than there were objectives with no difference, regardless of whether different or identical mastery standards were used.

Discussion

The district scores were clearly comparable with the state scores at the overall subject level for three of the four grades tested. There was strong evidence for the criterion-related validity of the subject scores yielded by the district tests for grades 3, 6, and 8, but there was only weak evidence for the criterion-related validity of those for grade 10. Consequently, the grade 10 tests and data should be examined further. In general, however, the subject scores yielded by the district tests can be regarded as substitutes for those yielded by the state tests.

With respect to objective mastery rates, the degree of comparability was low across all grades and subjects, indicating that objective mastery information obtained from the district test was not consistent with mastery information obtained from the state test. Even when identical mastery standards were applied to both batteries, the consistency of mastery classifications did not markedly improve. This result suggested that the discrepancies in the mastery rates between the two batteries were due to other factors besides mastery standards (e.g., item structure, item difficulty, content specifications).

When taken together, however, the data from this comparability study demonstrated support for the effectiveness of the model, both from the perspective of the state education agency and the local school district. The moderate to strong relationships between corresponding subject scores for the two batteries suggested that the district scores were, in general, valid measures of the state objectives. The weak relationships between mastery classifications implied that the district tests may be more specifically aligned with local curriculum than with state objectives. This was a positive implication, given that local autonomy in curriculum and assessment matters is valued by the state education agency. Moreover, the model was not intended to produce tests that were simply clones of the state tests, but rather to ensure that districts use accepted psychometric principles in adapting/developing tests to measure the state objectives.

Conclusions

The criteria for test approval featured in this model are based on the most important characteristic of test scores--validity. The principle that a criterion-referenced achievement test must demonstrate content validity is routinely applied in test evaluations. However, the collection of evidence pertaining to the criterion-related validity of scores adds an important dimension to the evaluation process. The statistical procedures utilized in the comparability study, when combined with careful scrutiny of the congruence ratings collected for the test under consideration, will ensure that any tests approved using this model will be valid measures of the state objectives.

The comparability data discussed in this paper illustrated how evidence for the criterion-related validity of district scores is collected. The finding that the district data demonstrated criterion-related validity but also seemed to reflect local curriculum and instructional practices suggested that the model serves the interests of both the state education agency and the district.

The unique three-stage approval process featured in this model allows a district to work toward approval with the benefit of specific feedback from the state education agency, thereby fostering

communication between the two agencies. The model also minimizes record keeping at the state education agency by putting the burden of documentation on the district applying for approval. By requiring districts rather than test publishers to apply for approval, the model promotes cooperation between test users and test publishers, ensuring that the consumer is purchasing what is needed. Furthermore, the two agencies involved in the approval process can manage the tasks required without either being overburdened.

In summmary, this validity-based model represents an effective mechanism for the evaluation of a criterion-referenced test. It allows a district to establish and maintain an autonomous testing program that is aligned with local curriculum and state objectives. It provides the state education agency with a way to ensure that both the spirit and the letter of the testing legislation is met. Thus, from both perspectives the model works.

References

American Psychological Association. (1985). Standards for educational and psychological testing. Washington, DC: Author.

Berk, R. A. (Ed.). (1980). Criterion-referenced measurement: The state of the art. Baltimore: The Johns Hopkins University Press.

Berk, R. A. (Ed.). (1980). A guide to criterion-referenced test construction. Baltimore: The Johns Hopkins University Press.

Hambleton, R. K. & Eignor, D. R. (1978). Guidelines for evaluating criterion-referenced tests and test manuals. Journal of Educational Measurement, 15, 321-327.

Madaus, G., Airasian, P., Hambleton, R., Consalvo, R., & Orlandi, L. (1979). Development and application of criteria for screening commercial standardized tests for the Massachusetts basic skills improvement policy. (Report prepared for the Massachusetts Department of Education). Boston: Public Affairs Research Institute.

Table 1

Sample Sizes for Comparability Study

	Subject			
Grade	Reading/English/Language Arts	Math	Science	Social Studies/Civics
3	373	356	408	394
6	404	410	419	428
8	296	393	314	287
10	385	353	355	324

Table 2

Validity Coefficients for Subject Scores

	Subject			
Grade	Reading/English/Language Arts	Math	Science	Social Studies/Civics
3	.792	.718	.724	.710
6	.813	.838	.828	.732
8	.741	.764	.778	.734
10	.444	.437	.514	.497

Table 3

Validity Coefficients for Objective Mastery/Nonmastery Classifications

Grade	No. of objectives	Phi coefficient		
		Median	Low	High
		Reading/English/Language Arts		
3	23	0.288	.156	.393
6	22	0.295	.138	.516
8	29	0.250	-.010	.461
10	29	0.249	.040	.426
		Mathematics		
3	17	0.330	.056	.538
6	26	0.319	.096	.518
8	25	0.290	.161	.492
10	23	0.330	.051	.443
		Science		
3	16	0.247	001	.385
6	23	0.236	.070	.442
8	18	0.195	-.007	.423
10	20	0.206	.066	.367
		Social Studies/Civics		
3	15	0.232	.025	.451
6	21	0.279	.079	.402
8	18	0.219	.124	.414
10	25	0.312	.149	.545

Table 4

Percentage of Objectives with Significantly Different Mastery Rates Using Different Mastery Standards

	Subject			
Grade	Reading/English/Language Arts	Math	Science	Social Studies/Civics
3	78%	65%	75%	87%
6	77%	85%	83%	76%
8	83%	92%	94%	83%
10	69%	87%	90%	68%

Table 5

Percentage of Objectives with Significantly Different Mastery Rates Using Identical Mastery Standards

	Subject			
Grade	Reading/English/Language Arts	Math	Science	Social Studies/Civics
3	70%	41%	75%	87%
6	73%	69%	87%	67%
8	76%	92%	94%	72%
10	72%	87%	80%	72%

Appendix A

Criteria for Approval of
District-Developed or Published Tests for Use in the
Measurement of the Key Skills

Introduction

The Excellence in Education Act of 1985 requires school districts to assess student performance on objectives or "key skills" identified by the Department of Elementary and Secondary Education (hereafter referred to as the Department) in English/reading/language arts, science, mathematics, and social studies/civics using criterion-referenced tests. These tests must be administered annually to students in at least two nonsequential levels within the grade span of two through six and to students in at least two nonsequential levels within the grade span of seven through ten.

A school district may use the Missouri Mastery and Achievement Tests (MMAT), develop its own tests, or use tests developed by a publisher, according to the provisions of the Excellence in Education Act of 1985. A district must obtain approval from the Department in order to use tests developed by the district or by a test publisher. This approval is contingent upon the district's demonstrating that certain criteria have been met.

The Criteria

The criteria a school district must meet are listed on the following pages. They are designed to insure that all tests used to assess performance on the key skills demonstrate desirable technical characteristics and yield results which can be used to improve instruction.

The criteria are based on the requirements set forth in the Excellence in Education Act of 1985. They are also based on the Standards for Educational and Psychological Testing, published in 1985 by the American Psychological Association, the American Research Association, and the National Council on Measurement in Education. These standards represent a summary of recommended practices in test construction and test use.

I. Test Development Plan and Rationale

Criterion: Any district constructing or purchasing tests to be used in assessing student performance on the key skills must submit to the Department a written plan that outlines the steps to be taken during the development of the the tests. The rationale for developing the tests and a list of the key skills that would be measured at each level within the grade span of two through ten must be included with this plan.

II. Validity

A. Key Skills Tests

Criterion: The tests must measure the key skills identified and tested by the Department at the grade levels selected for testing by the district in each of the following subjects: English/reading/language arts, science, mathematics, and social studies/ civics.

B. Content Specifications

Criterion: Each key skill measured by the tests must be expressed in the form of a content specification that clearly describes the learning outcome indicated by the key skill and how that outcome is to be measured. The content specification corresponding to each key skill must have been used in the selection or generation of items which measure that skill.

C. Content Representativeness

Criterion: Each item on the final test form(s) must have been judged to be a measure of (be congruent with) the key skill it is intended to assess, thereby facilitating alignment between each key skill, its test content specification, and the items measuring that skill. This should be done using a recognized and appropriate methodology, such as obtaining item-key skill congruency ratings from at least three content experts. In a congruency rating analysis, an item should receive a median rating of at least two on a five-point scale (1 = Excellent) in order to be considered congruent with its key skill.

D. Comparability to State-Developed Tests

Criterion: Statistics must be computed which relate the results yielded by the tests being submitted for approval to those yielded by the Missouri Mastery and Achievement Tests. The Department will compute these statistics using data collected by the district for key skills and subjects that are measured at the same grade levels by both tests. This will be done by dually administering the Missouri Mastery and Achievement Tests and the district-developed or published tests to a representative sample of students.

III. Scores

A. Individual Examinee Performance

Criterion: An examinee must be designated as a master or nonmaster of each key skill; these designations must be reported in an easily understood format. (In order to compute reliability estimates and utilize standard errors of measurement, a numerical score scale for each key skill must be determined.)

B. Determination of Mastery

Criterion: A recognized and appropriate standard-setting methodology must have been used in the determination of each mastery score, such as a judgmental or empirical-judgmental procedure. The department will assist with standard-setting procedures upon request.

C. Group Performance Data

Criterion: It must be possible for a district to analyze and report group performance data by grade level, building, and district for each key skill and subject.

D. Interpretation of Results

Criterion: Information must be provided to students, teachers, administrators and parents in an easily understood format regarding interpretation of individual and group performance. Suggestions for using test results to improve instruction and student learning must also be offered to test users.

IV. Reliability

A. Test Length

Criterion: Each key skill must be measured by at least four congruent items.

B. Score Reliability

Criterion: Score reliability estimates and standard errors of measurement must be determined for each key skill and for each subject. The Department will assist with the computation of these data upon request.

C. Mastery/Nonmaster Classification Consistency

Criterion: An index of consistency indicating the reliability of mastery/nonmastery classifications for each key skill must be determined. The Department will assist with the computation of these indices upon request.

V. Bias

Criterion: The "research on the influence of cultural diversity on test performance" (The Excellence in Education Act of 1985, Section 4.3) must have been considered during test development. A recognized and appropriate judgmental and/or statistical methodology must have been used to identify and reduce bias and stereotyping.

Appendix B

Application for Test Approval

Part I

Please enter the following:

NAME OF DISTRICT: _____

STREET ADDRESS: _____

CITY/STATE/ZIP: _____

DATE: _____

CONTACT PERSON: _____

TELEPHONE: _____

Are new tests being developed? _____

If so, what will they be called? _____

Is a published battery being adapted? _____

If so, what battery and by whom is it published?

_____

Use the guidelines included with this application to complete these three steps.

Attach a statement describing your rationale for developing a battery. The guidelines include questions which will facilitate developing a rationale statement.

Attach a list of the key skills that would be measured at each level within the grade span of two through ten. This list should also indicate the four (or more) grade levels selected for testing. The guidelines describe requirements for this list.

Answer the questions included in the guidelines in order to develop a detailed plan for the construction or adaptation of the battery. These questions were designed to facilitate the development of a plan which will result in a battery that meets the approval criteria.

References pertaining to the construction of criterion-referenced tests are available from the Department of Elementary and Secondary Education upon request. Department personnel are available for technical assistance with the development of your plan or at any other time during the construction and approval process.

Be sure to return this form, including your rationale, key skills list and test development plan by October 15 to:

Testing and Assessment Section
Department of Elementary and Secondary Education
P.O. Box 480
Jefferson City, MO 65102
(314) 751-3545

Appendix C

Application for Test Approval

Part II

NAME OF DISTRICT: _____

DATE OF FILING: _____

Respond to each statement by checking the appropriate response to indicate whether certain steps were taken and necessary data collected. Enter information that is requested in the indicated spaces. Use the accompanying worksheet and data forms to complete this part.

Validation Steps and Data

___ A. test measures key skills currently identified

___ B. items derived from test content specifications

___ C. items judged to be congruent

___ D. data for comparability study obtained; completed Data Form 5 attached.

Mastery Scores and Score Reporting

___ A. a mastery score determined via acceptable methodology for each curriculum area test and each key skill test

___ B. examinee score reporting format and interpretive guides prepared

___ C. group data summary formats and interpretive guides prepared

Reliability Studies and Data

___ A. at least four congruent, unique items per skill

___ B. an estimate of score reliability and a standard error of measurement for each curriculum area test and each key skill test computed

___ C. an index of consistency for each key skill mastery/nonmastery classification computed

(over)

Bias (only one technique needs to be utilized)

___ A. items judged to be free of bias and stereotyping by committee

List name of each member of Bias Review Committee and group each represents.

___ B. items determined to be free of bias via statistical procedure

Describe statistical procedure used to detect item bias.

_____ All necessary data is on file and ready for DESE review.

(If necessary data not on file, explain exceptions below.)

Signature of Test Development Director: _____

Return this form to:

Testing and Assessment Section
Department of Elementary and Secondary Education
P.O. Box 480
Jefferson City, MO 65102
(314) 751-3545

If the above responses indicate the appropriate data have been obtained, the Department will contact the district concerning a review of the data forms and the worksheet.